Gender-Driven Emotion Recognition System Using Speech Signals For Human Computer Intelligent Interaction

Mekhala SriDevi Sameera, A Satish Kumar, Kotte sandeep

Abstract -**This paper proposes a peculiar and very important developing area concerns the remote monitoring of elderly or ill people. Indeed, due to the increasing aged population, Human-Computer Intelligent Interaction (HCII) systems able to help live independently are regarded as useful tools. In this context recognizing people emotional state and giving a suitable feedback may play a crucial role. The purpose of speech emotion recognition system is to automatically classify speaker's utterances into seven emotional states including anger, boredom, disgust, fear, happiness, sadness and neutral state. Emotions have been classified separately for male and female based on the fact male and female voice has altogether different range. It provides a solution by improving interaction among human and computers, thus allowing human-computer intelligent interaction. The system is composed of two subsystems: 1) gender recognition (GR) and 2) emotion recognition (ER). It distinguishes a single emotion versus all other possible ones as in proposed numerical results. Speech based emotion recognition system consists of four principle parts: Feature Extraction, Feature Selection, Database and Classification.**

Nowadays, the research is focused on finding powerful combinations of classifiers that increases the classification efficiency in real-life speech emotion recognition applications. From these acoustic signals, this project will calculate pitch, short time energy, zero crossing rate and Mel frequency cepstral coefficients, and correlate it to emotions of the driver. We also define these features and the feature extraction methods. In paper, a demonstration on how one can distinguish the emotion based on these features (or combination of features) by testing them over Berlin emotion database

Index Terms—**Classification, Emotion Recognition, Feature Extraction, Human Computer Interaction**

I. Introduction

Human emotion recognition is an important element for proficient Human Computer Intelligent Interaction. There has been a growing interest to improve human-computer interaction. It is well-known that, to achieve efficient Human-Computer Intelligent Interaction (HCII), computers should be able to interact naturally with the users, i.e. the mentioned interaction should ape human-human interactions.

In this context recognizing people's emotional state and giving a suitable feedback may play a crucial role. HCII is becoming really significant in applications such as smart home, smart office and virtual reality, and it may obtain importance in all aspects of future people's life. Indeed, due to the increasing aged population, HCII systems able to help live independently are regarded as useful tools. Despite the fact of extensive advances intended at supporting elderly citizens, many issues have to be addressed in order to help aged ill people to live independently. The system is composed of two subsystems: 1) gender recognition (GR) and 2) emotion recognition (ER).

An emotion is a mental and a physiological state associated with a wide variety of feelings, thoughts, actions and behavior. The system is able to recognize six emotions (anger, boredom, disgust, fear, happiness, and sadness) and the neutral state. This set of emotional states is typically used for emotion recognition purposes. It also distinguishes a single emotion versus all the other possible ones, as proven in the proposed numerical results. Emotions have been classified individually for male and female based on the fact male and female voice has overall different range. It provides a solution by improving interaction among human and computers, thus allowing human-computer intelligent interaction. As a consequence, emotion recognition represents a hot research area in both industry and academic field. Generally emotion recognition systems are based on facial or voice features.

When we analyze audio signals, we usually opt the method of short-term analysis since most audio signals are more or less secure within a short period of time, say 20 ms. When we do frame blocking, there may be some overlaps between neighbouring frames to confine slight change in the audio signals. Within each frame, we can observe the three most diverse acoustic features, for human as follows:

Manuscript received May 16, 2015.

Mekhala Sridevi Sameera, Computer Science and Engineering, Dhanekula Institute of Engineering & Technology, Vijayawada, India,

A Satish Kumar, Computer Science and Engineering, Dhanekula Institute of Engineering & Technology, Vijayawada, India,

Kotte Sandeep, Computer Science and Engineering, Dhanekula Institute of Engineering & Technology, Vijayawada, India,

- Volume: It correlates to the compression of your lungs. A large compression in the lungs corresponds to a large volume of audio signals.
- Pitch: It correlates to the vibration frequency of your vocal cord. A high pitch corresponds to a high vibration frequency.
- Timbre: It correlates to the positions and shapes of your lips and tongue. A different timbre corresponds to different positions and shapes of your lips and tongue

This paper proposes a solution, intended to be employed in a Smart Environment, able to confine the emotional state of a person starting from a registration of the speech signals in the surrounding obtained by mobile devices such as smart phones.

II. Feature Extraction & classification

Feature extraction is used to reduce the large input data into smaller data and it converts the data into small feature sets or feature vectors (n-dimensional vector to store numerical features which represents an object). Feature extraction is defined as the process of extracting the feature from a source data, where the data can be embedded from high dimensional data set [1]. We calculate different feature sets for different applications. For example in computer vision applications edges and corners are calculated as features for images, features like noise ratio, length of sound and relative power are calculated for pattern recognition applications. Here, in this work the input data is taken from microphone and the feature extraction algorithms are executed to extract these features in real time. Features are extracted from the real time data by performing time and frequency domains algorithms. These algorithms extract temporal features, spectral features and cepstral coefficients. These features are extracted based on the amplitude and spectrum analyzer of the audio data. Figure.1 explains how the audio features are extracted from the input data. The process starts with dividing into sequence of frames which is called as windowing. Then we perform the feature extraction methods for estimating acoustic features that are mostly used in emotion detection. Pitch, zero-crossing rate, short time energy and MFCC are used to extract acoustic features. Zero-crossing rate and short time energy are calculated for the voiced and unvoiced signals. After adding energy, delta, and double delta features to the 12 cepstral features, totally 39 MFCC features are extracted.

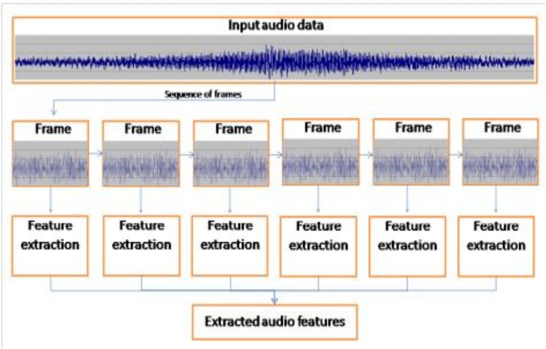

Figure 1: Feature extraction process

Monitoring the emotions are favourable based on non intrusive techniques, since these does not results much in stress, annoyance etc. So we are focusing on real time speech signals. The real time emotion detection system performance depends upon the given input speech signals. The general problems identified in this type of system are as follows:

Converting the stream of audio signals into set of frames by performing frame blocking. Catching of the time-varying characteristics of the audio signals cannot be done if the frame duration is too big, so the time duration of each frame is about 20 to 30 ms. But to extract valid acoustic features, the frame duration should not be too small, then extraction cannot be done.

In the window operation, the large input data is divided into small data sets and stored in sequence of frames. While dividing some of the input data may be discontinuous. To achieve the continuity the sequence of frames of the input data are overlapped. This window operation is performed using hamming window method to reduce the spectral leakage in the input data.

The resulting feature vector usually has very large dimension, which leads to unsatisfactory performance of emotion detection system. In order to select good features, we should improve accuracy and reduce the dimensions of feature vector.

To overcome these performance problems linear subspace techniques are used. Linear subspace techniques are statistical related techniques which are used to reduce the dimensionality and to classify the given data.The linear subspace techniques are successful to some extent but they are not effective in large databases [5][4][6]. LDA uses information about the class [3]. It tries to maximize the between class variance and minimize the within class variance. In other words, it decreases the distance between same class files and increases the distance between different class files [2]. Because of that LDA easily recognizes the emotions among large databases.

The LDA performs considerably better when compared with above classifiers for large databases. The performance of LDA can also be increased if we replace the linear metric with non linear metric as explained below.

1. A set of MFCC acoustic vectors result from the initial voice signal. Each of them is composed of 256 samples, but the speech information is codified mainly in the first

12 coefficients. Therefore, each acoustic vector is truncated at its first 12 samples and then it is positioned as a column of a matrix.

2. The resulted MFCC acoustic matrix constitutes a powerful speech discriminator which works successfully as a feature vector for the processed signal.

3. Each feature vector has 12 rows and a number of columns depending on the length of the speech signal. Therefore, because of their different dimensions, these speech feature vectors cannot be compared using linear metrics, such as the most known Euclidean distance.

4. An alternative solution with Euclidean distance can be done through re-sampling or padding with zero values so that they get same dimensions. But the disadvantage of this approach is possible loss of speech information from the feature vectors.

5. For this reason, a special nonlinear metric such as Hausdorff distance metric is introduced which is able to compute the distance between different sized matrices having a single common dimension [8][7].

6. The success rate is increased to 85% approximately.

The LDA approaches are depending upon data set classifications [9].

In class specific method, we maximize the ratio of the between class scatter matrix to the within class scatter matrix for each class separately. The class separation is obtained by using this ratio. Two optimization criterions are used to transform the data sets independently.

In class independent method, we maximize the ratio of overall variance to within class variance or maximizing the between class scatter to within class scatter matrix across all classes simultaneously. In class independent method, one optimization criterion is used to transform the data sets.

III. BERLIN DATABASE

The Berlin database contains speech with acted emotions. Emotion is an important factor in communication. It is developed by the Technical University, Institute for Speech and Communication, Department of Communication Science, Berlin [10]. It has become one of the most conventional databases used by researchers on speech emotion recognition, thus facilitate concert comparisons with other studies. There are 5 actors and 5 actresses who contributed speech samples for this database, mainly have 10 German utterances, 5 short utterances and 5 longer ones and are recorded with 6 kinds of emotions: anger, boredom, disgust, fear, happiness, sadness.

IV. CONCLUSION

For the machine based state estimation, we will not focus on the voice content but rather on voice-signal features that are relevant for an emotional state inference. In this regard, to make the system more robust in predicting the emotion state we analyzed the acoustic information such as pitch, short time energy, zero crossing rate and MFCC etc. in order to extract appropriate features. From the literature emotion recognition based on acoustic information has been implemented on a variety of classifiers. Tentative results show that the LDA classifier with linear metric produces 67.22% recognition rate. To improve the recognition rate later we used a special non linear metric called Hausdorff distance measure. The recognition rate is improved to 85% approximately.

REFERENCES

[1] I.M. Guyon, S.R. Gunn, M. Nikravesh, and L. Zadeh, editors. Feature Extraction, Foundations and Applications. Springer, 2006.

[2] Hua Ai, Diane J. Litman, Kate Forbes-riley, Mihai Rotaru, Joel Tetreault, and Amruta Pur. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In Proceedings of Interspeech, pages 797-800, 2006.

[3] Laurence Devillers and Laurence Vidrascu. Real-life emotion detection with lexical and paralinguistic cues on Human-Human call center dialogs. Proc. INTERSPEECH' 06. Pittsburgh, 2006.

[4] J I. Murray, Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America, 93(2):1097-1108, 1993.

[5] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. Speech Communication, 48(9):1162 -1181, 2006.

[6] Thurid Vogt, Elisabeth Andre, and Johannes Wagner. Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realization. pages 75-91, 2008.

[7] T. Barbu. Discrete speech recognition using a hausdorff-based metric. In In Proceedings of the 1st Int. Conference of E-Business and Telecommunication Networks, ICETE 2004, volume 3, pages 363_368, Setubal, Portugal, Aug 2004.

[8] T. Barbu. Speech-dependent voice recognition system using a nonlinear metric. In International Journal of Applied Mathematics, volume 18, pages 501-514, 2005.

[9] J.Beveridge K.Baek, B. A.Draper and K. She. Analysis of pca-based and fisher discriminant-based image recognition algorithms. Technical report, Department of Computer Science., 2000.

[10] M. Rolfes W. Sendlmeier F. Burkhardt, A. Paeschke and B. Weiss. Berlin database of emotional speech on-line. In Interspeech: http://pascal.kgw.tu-berlin.de/emodb/index-1024.html, pages 1517_1520, 2005

Mekhala Sridevi Sameera pursuing her M.Tech in Computer Science and Engineering from Dhanekula Institute of Engineering & Technology affiliated to JNTUK. Graduated in Computer Science and Engineering from Dhanekula Institute of Engineering & Technology in 2013.

A.Satish Kumar working as Assistant professor in the department of Computer Science and Engineering in Dhanekula Institute of Engineering & Technology, India. Received his M.Tech in Computer Science and Engineering from sathyabama university, Chennai.

Kotte Sandeep working as Assistant professor in the department of Computer Science and Engineering in Dhanekula Institute of Engineering & Technology, India. Graduated in Information technology from JNTUH in 2007 and M.S in Information technology from the University of Klagenfurt, Austria in 2010 specialized in Intelligent Transportation System, pervasive computing and Business Informatics.